Bioinformatics for PIs:

Bioinformatics Workflows

Agata Smialowska

NBIS
Stockholm, 17 September 2019

agata.smialowska@nbis.se

NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

UNIVERSITET STOCKHOLMS

In the internet browser:

https://pollev.com

agatas031
(not case-sensitive)

To show this poll

1

Install the app from pollev.com/app

2

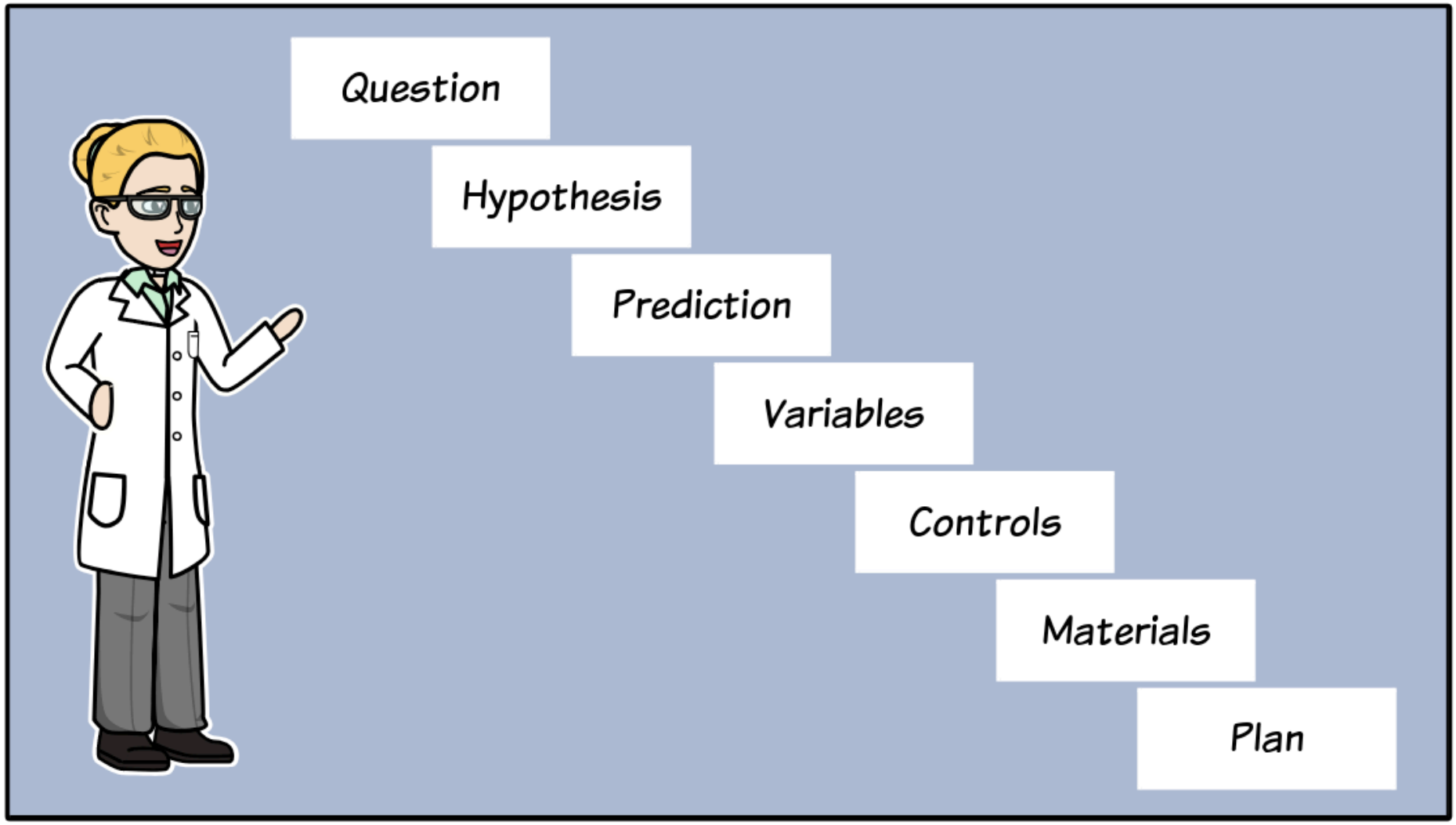
Start the presentation

Still not working? Get help at pollev.com/app/help

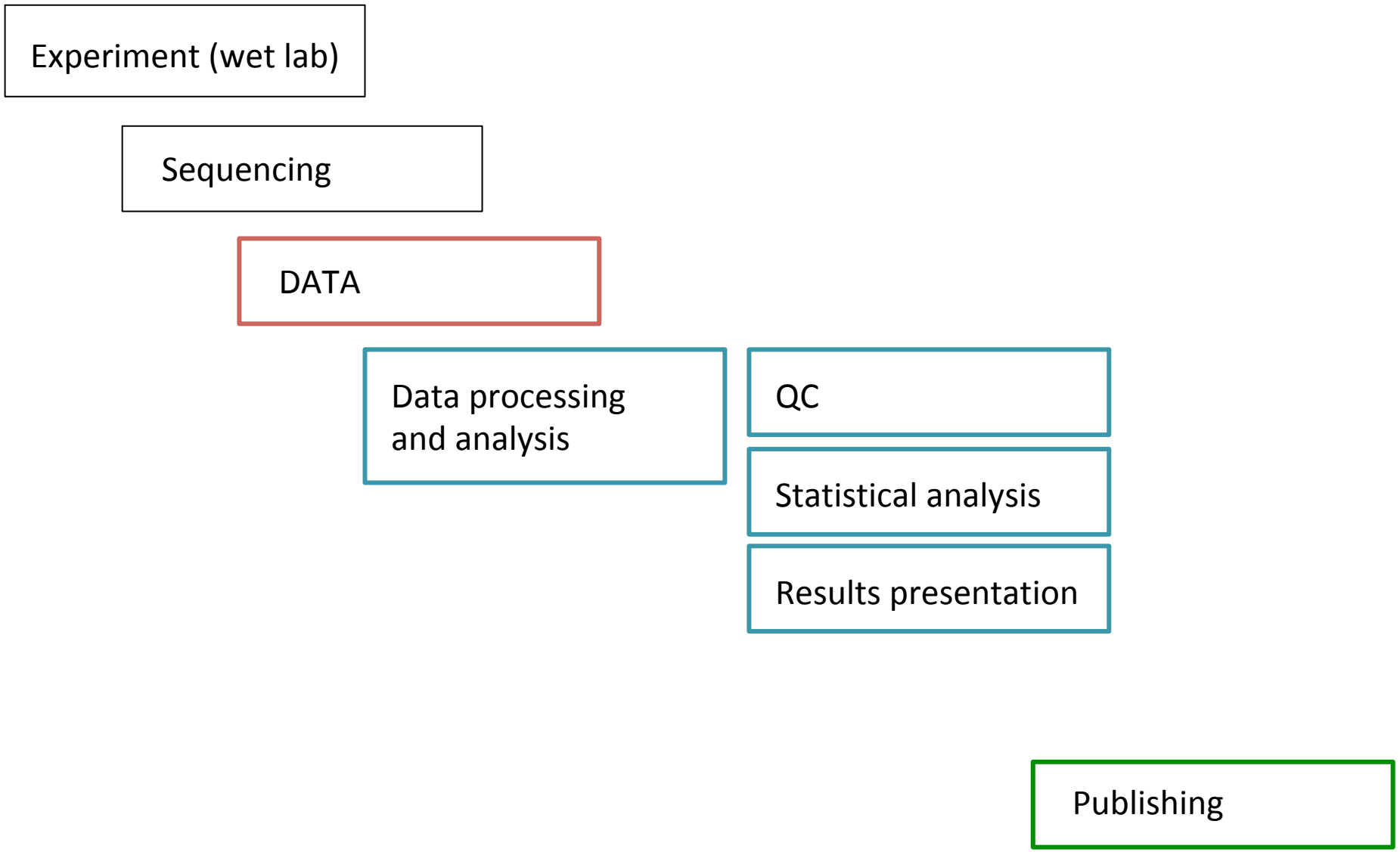
or

Open poll in your web browser

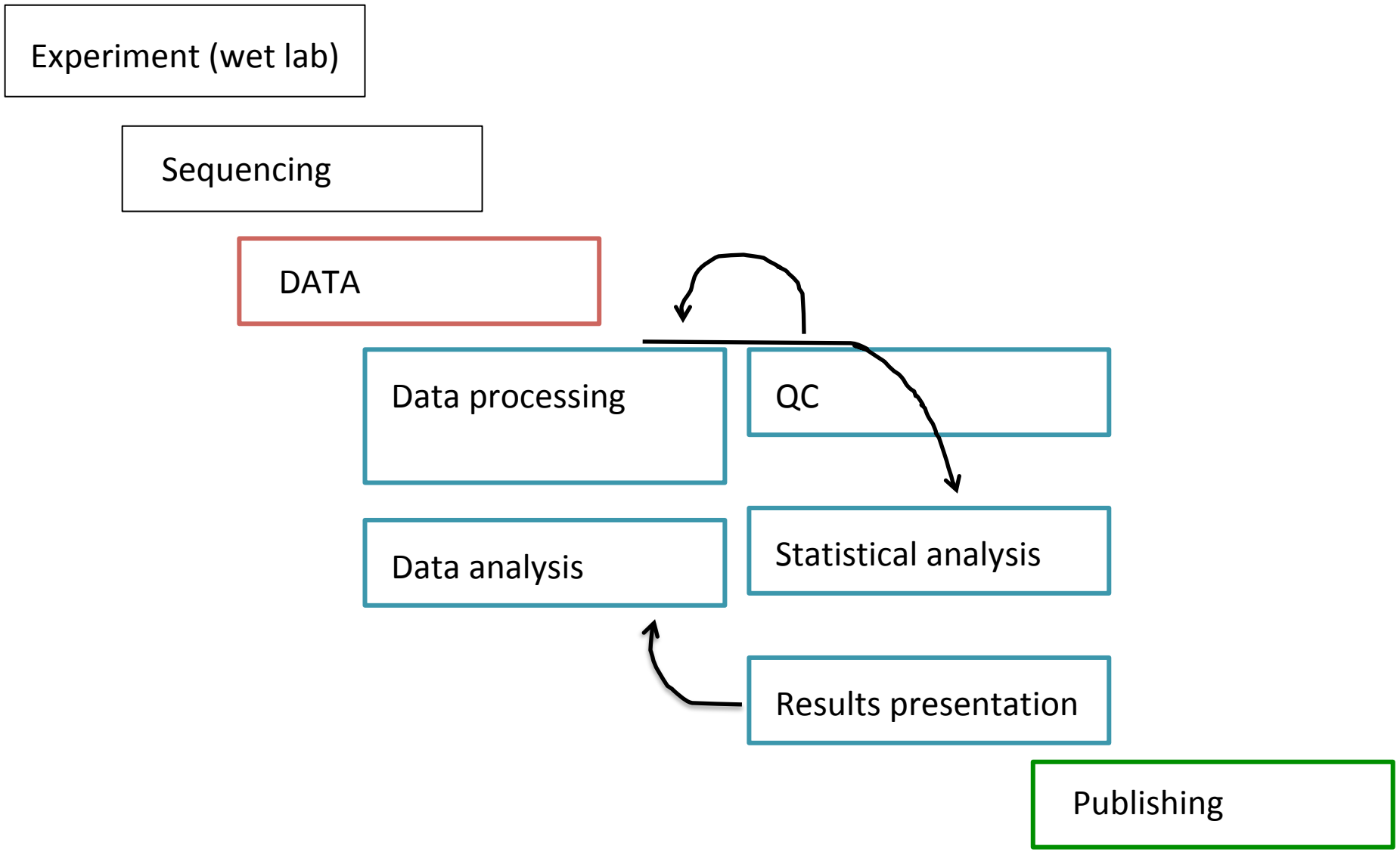
Life cycle of a scientific project – part 1



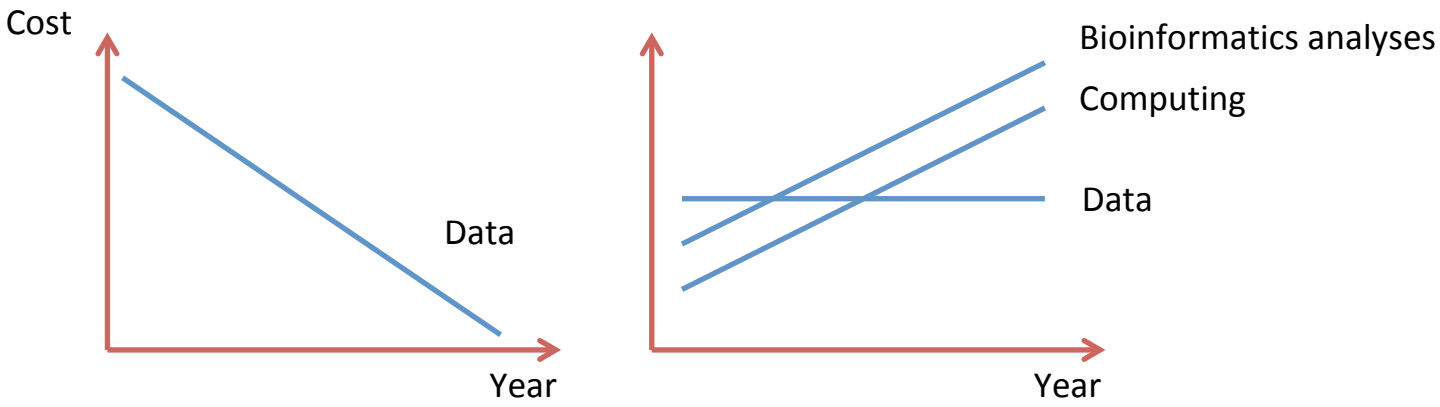
Life cycle of a scientific project – part 2



Life cycle of a scientific project – part 2



Production is cheap, analysis is not

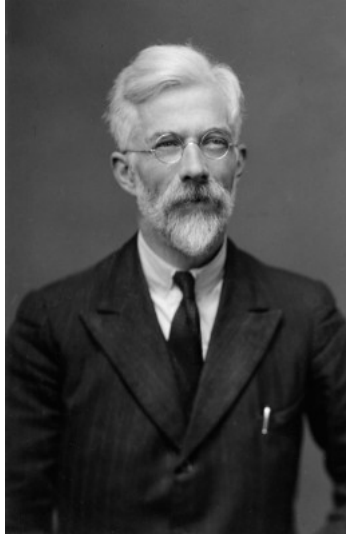


"Per base" "Per project"

Concepts for data-driven research

- reproducible research
- FAIR data
- Integrative omics

Experimental design



A simple truth:
There is no technology nor statistical wizardry that can save a poorly planned experiment. The only truly failed experiment is a poorly planned one.

To consult the statistician after an experiment is finished is often merely to ask them to conduct a post mortem examination. They can perhaps say what the experiment died of. (Ronald Fisher, 1938)

Experimental design

- Sound experimental design: replication, randomisation and blocking (R. Fisher, 1935)
- In the absence of a proper design, it is essentially impossible to partition biological variation from technical variation

- <u>To think about:</u>
  - Batches: Design your experiment to avoid *confounding* your different treatments (sex, nutrition) with each other or with technical variables (lane within a flow cell, between flow cell variation)
  - Statistical power to detect differences of interest (effect size, number of biological replicates)

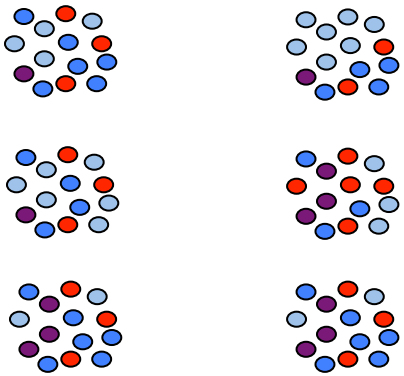
Experimental design

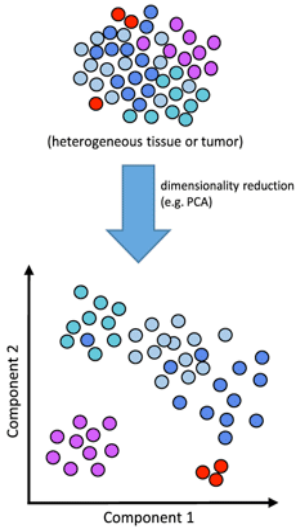
We encourage you to discuss the experimental design with the person who will analyse the data (or with NBIS)!

Formulating the scientific question and means to address it

Design and Technical aspects

Choice of experimental setup





Low-input bulk RNA-seq

Single cell RNA-seq

To show this poll

1

Install the app from pollev.com/app

2

Start the presentation

Still not working? Get help at pollev.com/app/help

or

Open poll in your web browser

1. Introduction to workflows frequently used in bioinformatics of NGS data
2. Workflows for transcriptomics
    1. Bulk RNA-seq
    2. Small RNA-seq
    3. Variant discovery in RNA-seq data
3. Workflows for functional genomics
    1. ChIP-seq (TF)
4. Comments

1. Introduction to workflows frequently used in bioinformatics of NGS data
2. Workflows for transcriptomics
    1. Bulk RNA-seq
    2. Small RNA-seq
    3. Variant discovery in RNA-seq data
3. Workflows for functional genomics
    1. ChIP-seq (TF)
4. Comments

Workflow for NGS data processing and analysis

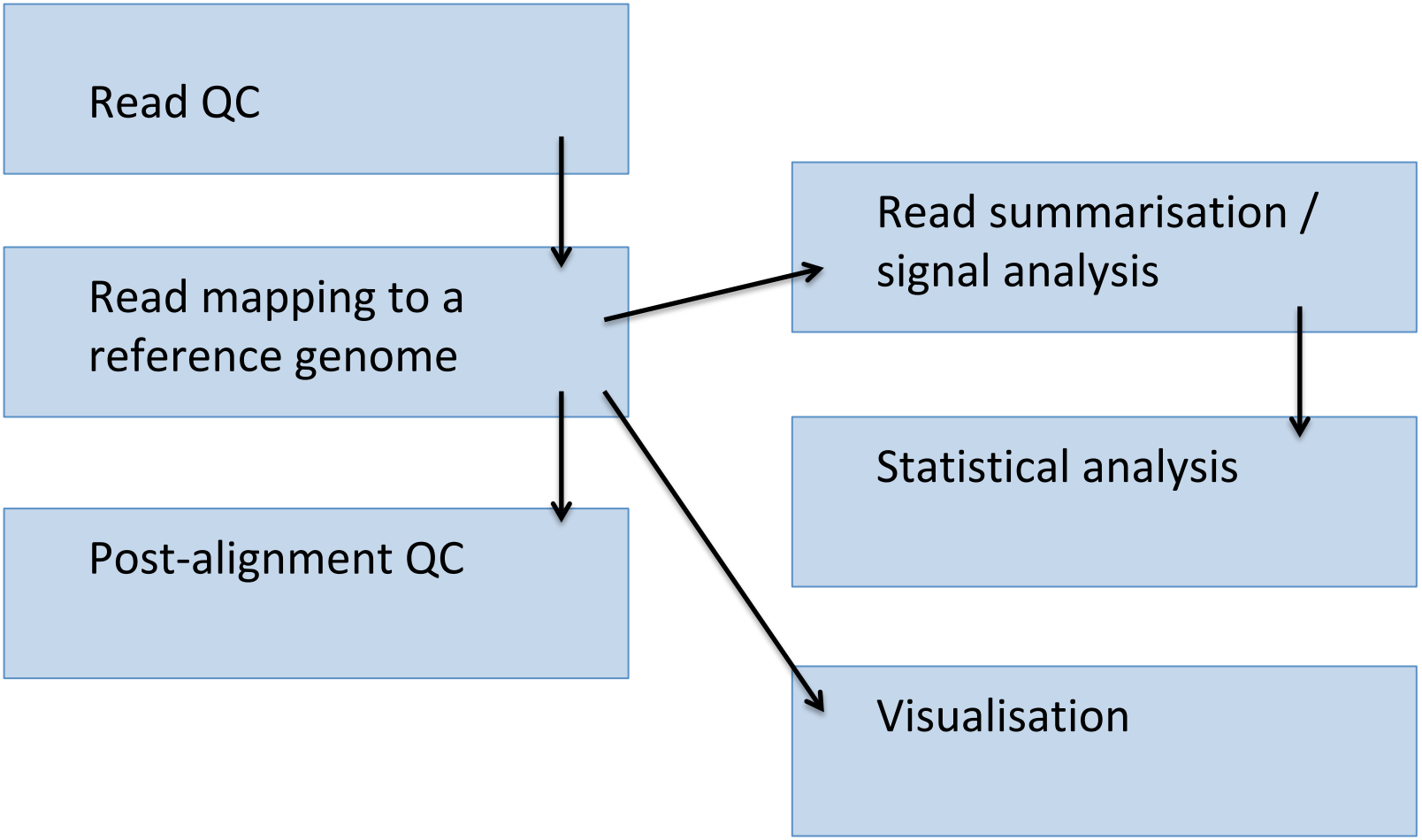
- Which reference genome? (newest assembly usually best) From which source (Ensembl, UCSC)

- Which read mapping strategy (global, local, alignment reporting, multimapping reads, which SAM tags to include, etc)? Which mapper?

- Post-alignment processing? Strategy?

- Read summarisation strategy? (reads with unique best alignments vs. all mapped reads, count all occurrences or in fractions?)

Workflow for NGS data processing and analysis



Workflow for NGS data processing and analysis

- Which reference genome? (newest assembly usually best) From which source (Ensembl, UCSC)

- Which read mapping strategy (global, local, alignment reporting, multimapping reads, which SAM tags to include, etc)? Which mapper?

- Post-alignment processing? Strategy?

- Read summarisation strategy? (reads with unique best alignments vs. all mapped reads, count all occurrences or in fractions?)

Workflow for NGS data processing and analysis

- It may be tempting to use an available solution (there are many pipelines out there)...

Workflow for NGS data processing and analysis

- It may be tempting to use an available solution (there are many pipelines out there)...

- **No "One size fits all" solutions available**

Workflow for NGS data processing and analysis

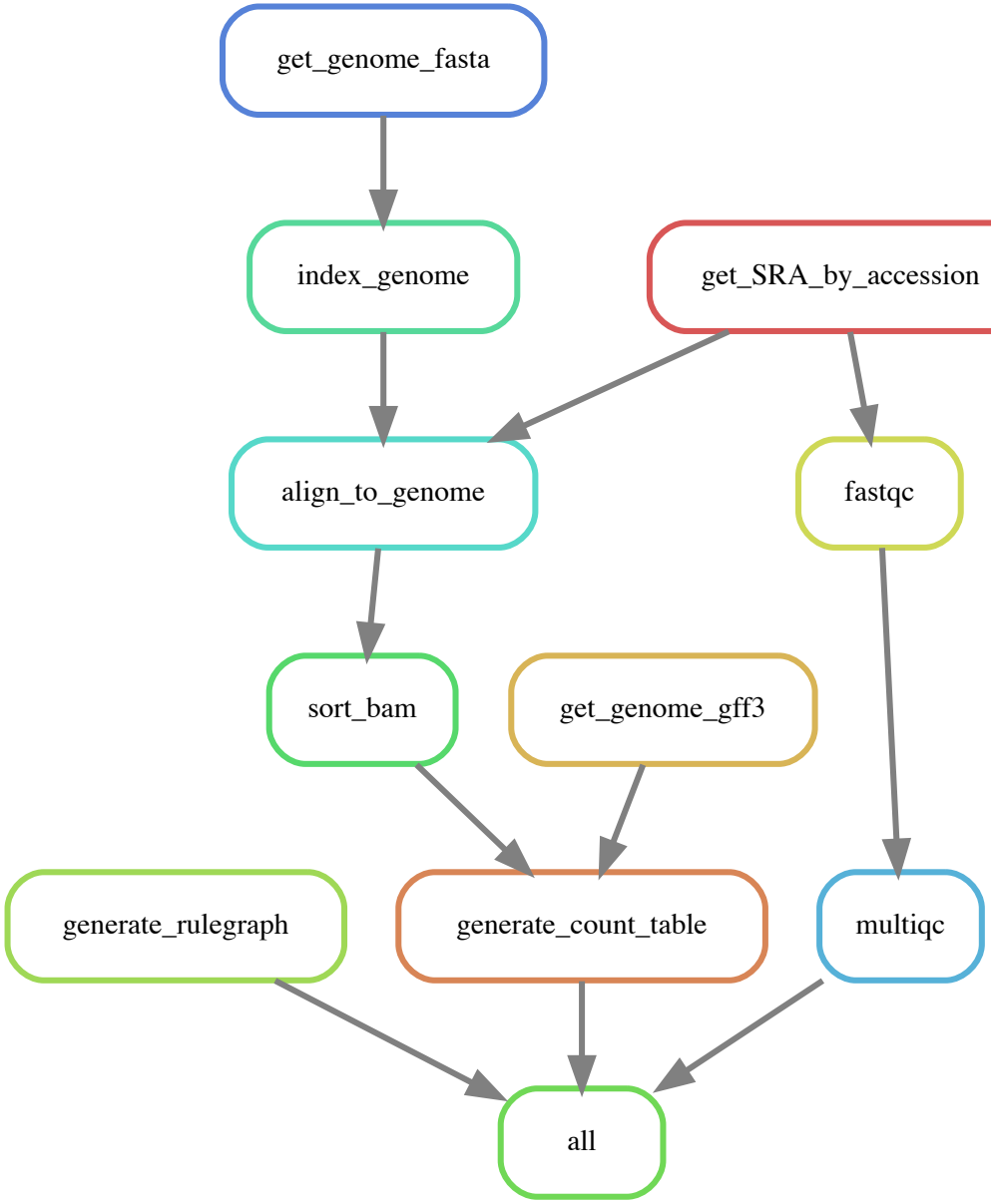
- It may be tempting to use an available solution (there are many pipelines out there)...

- **No "One size fits all" solutions available**

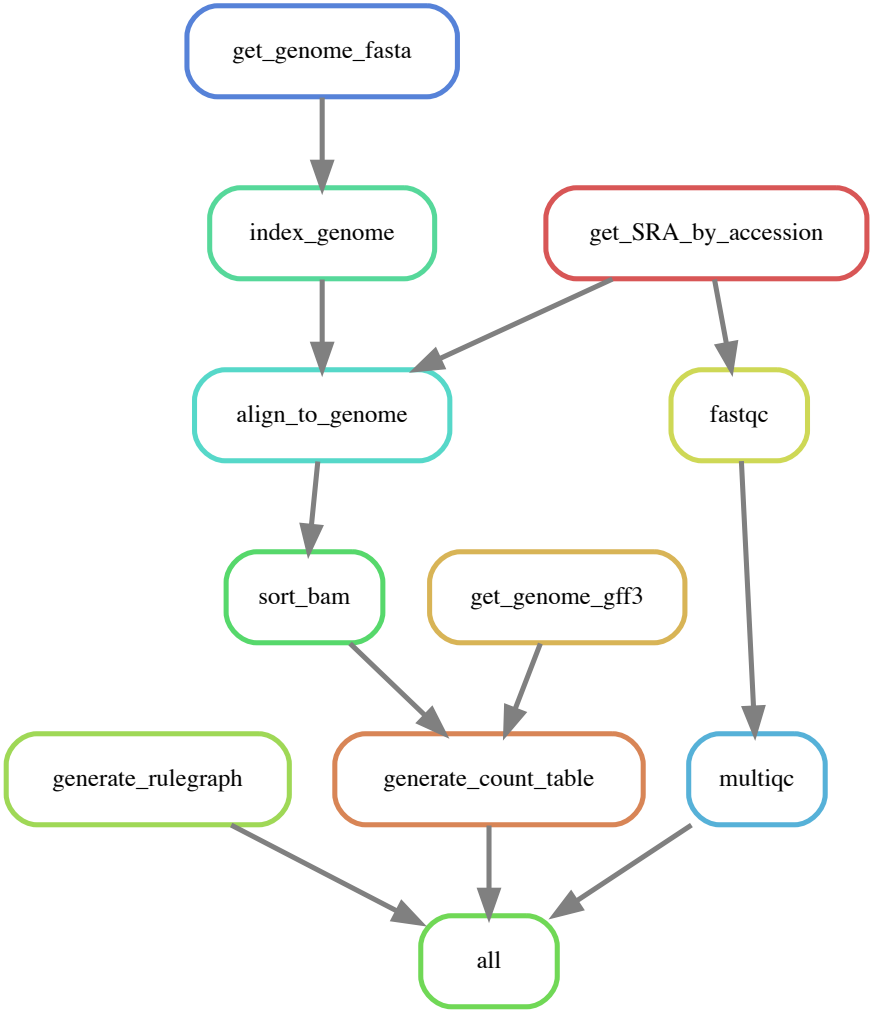
- **Best Practice** solutions (for standard cases):
  - **ENCODE** for functional genomics
  - **GATK** for variant analysis
  - Many resources for RNA-seq: Bioconductor, review publications, NBIS (!) – depends on the actual question behind the experiment
  - NGI maintains pipelines for most of standard applications

1. Introduction to workflows frequently used in bioinformatics of NGS data
2. Workflows for transcriptomics
    1. Bulk RNA-seq
    2. Small RNA-seq
    3. Variant discovery in RNA-seq data
3. Workflows for functional genomics
    1. ChIP-seq (TF)
4. Comments

Bulk RNA-seq (differential expression)

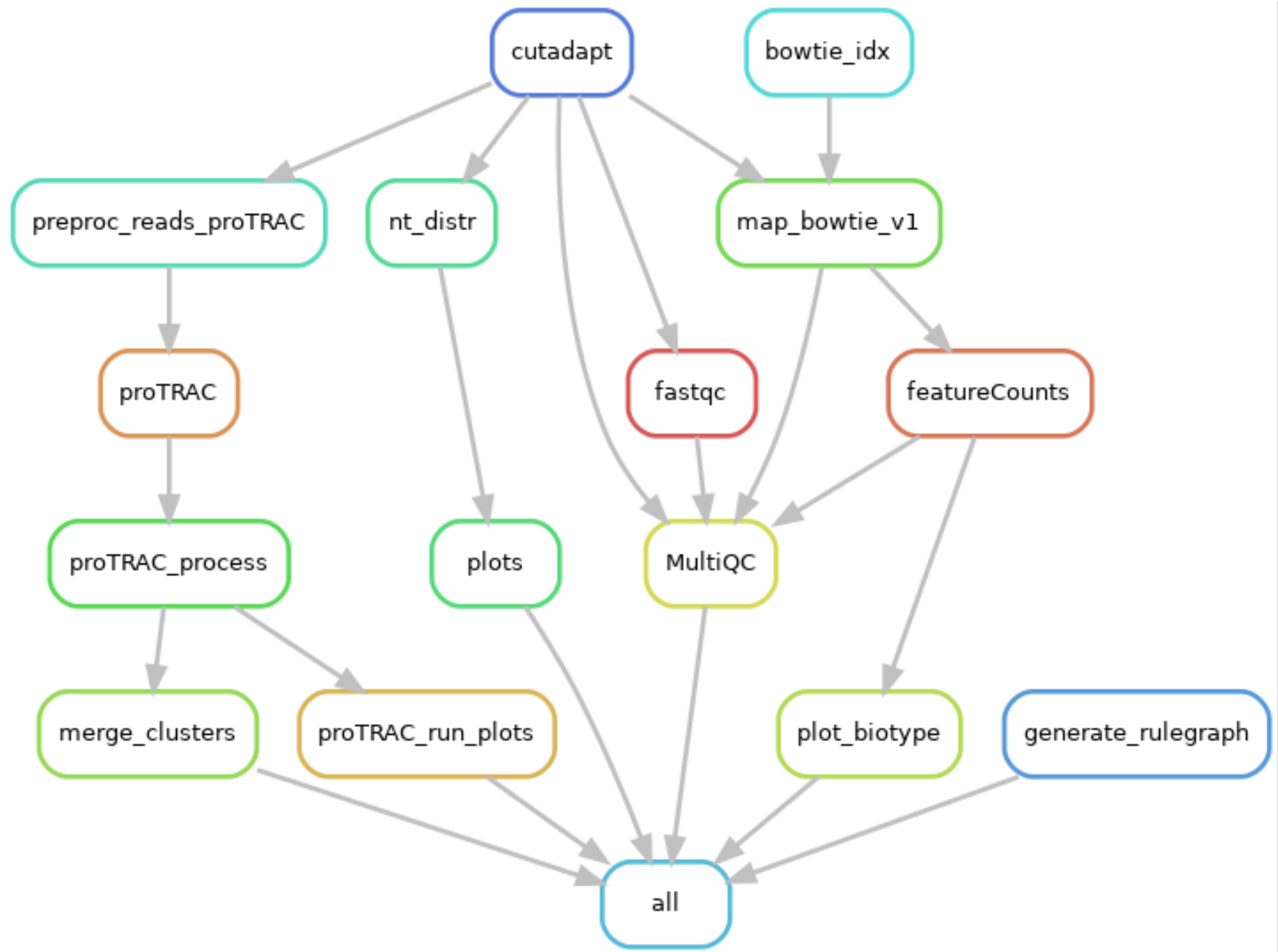


Bulk RNA-seq (differential expression)

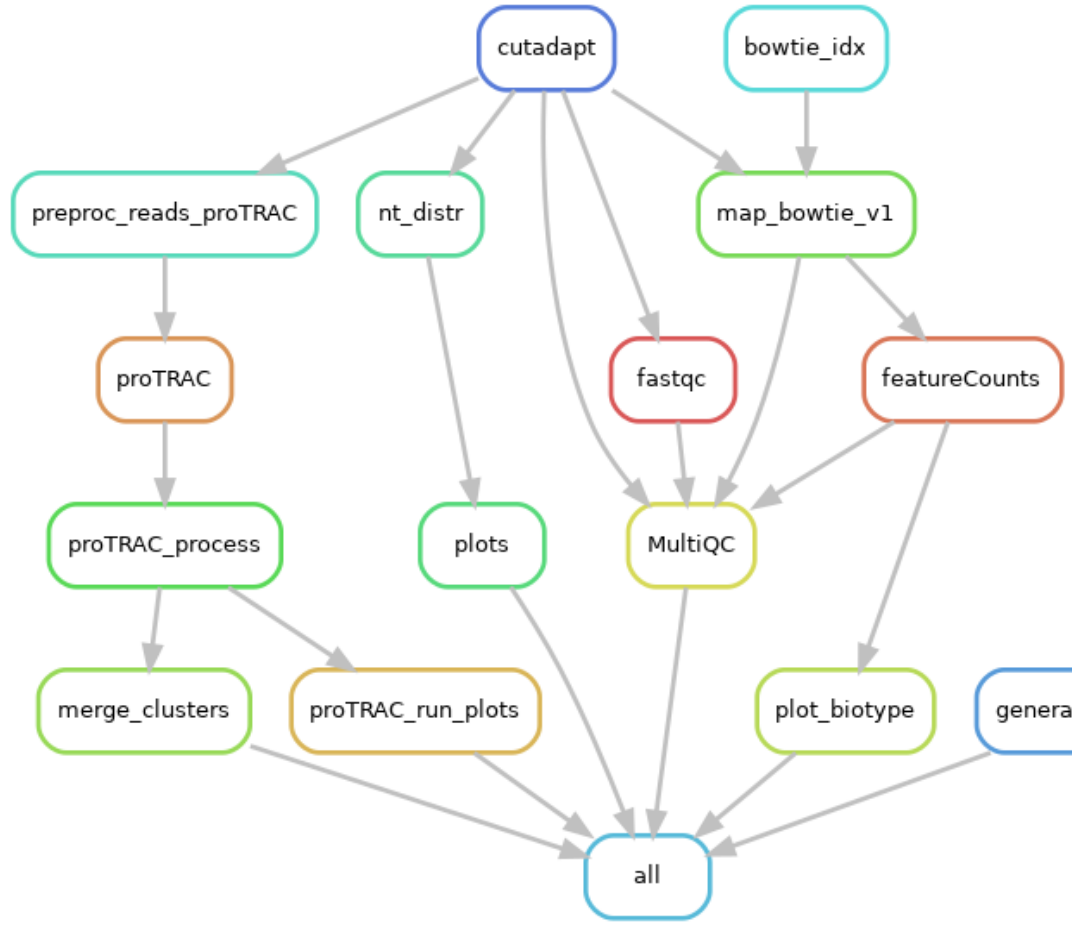


- Generation of reference genome index
- Get fastq files from SRA
- Read QC
- Mapping to reference genome
- Read counting
- Aggregation of QC metrics and program logs

Small RNA-seq (piRNA)

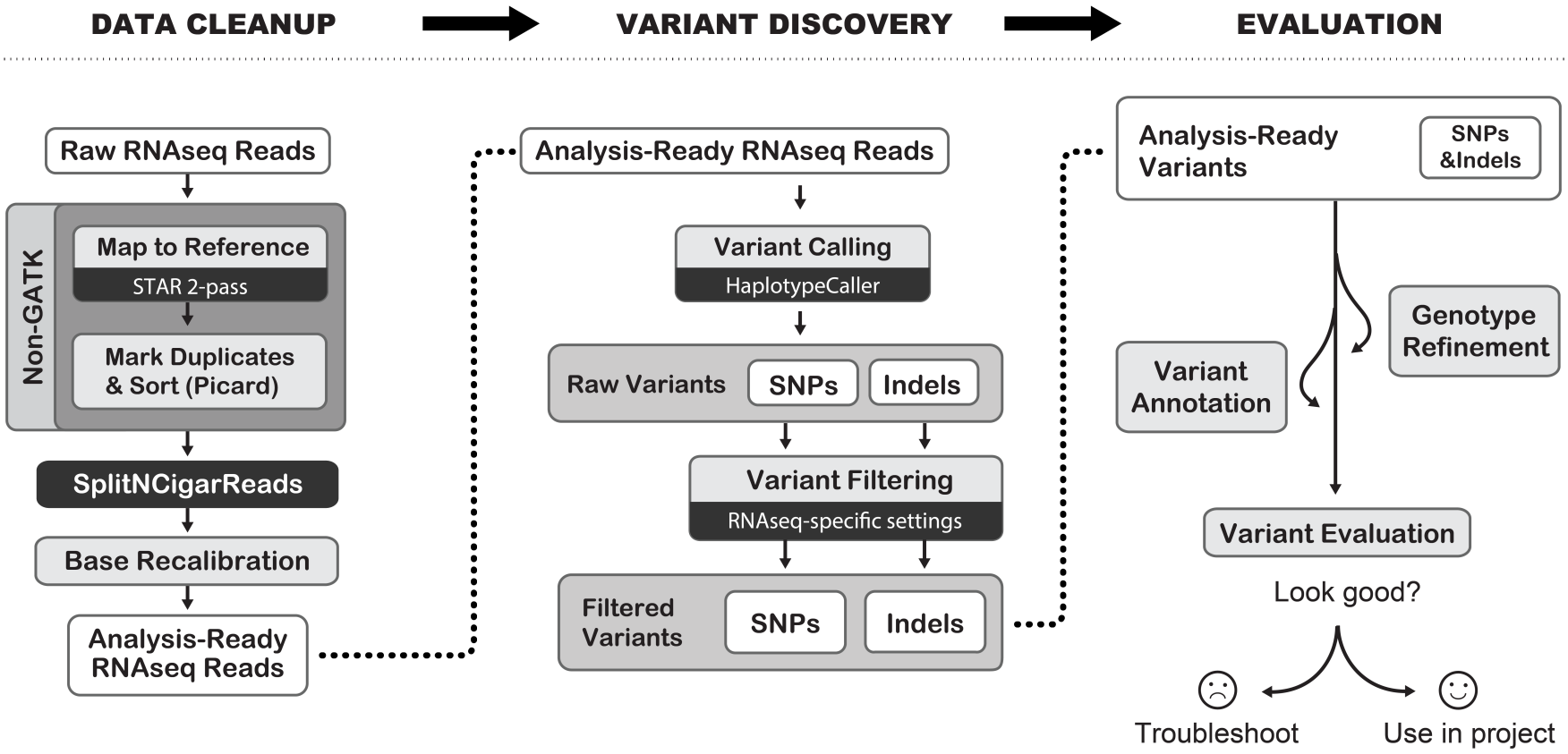


Small RNA-seq (piRNA)



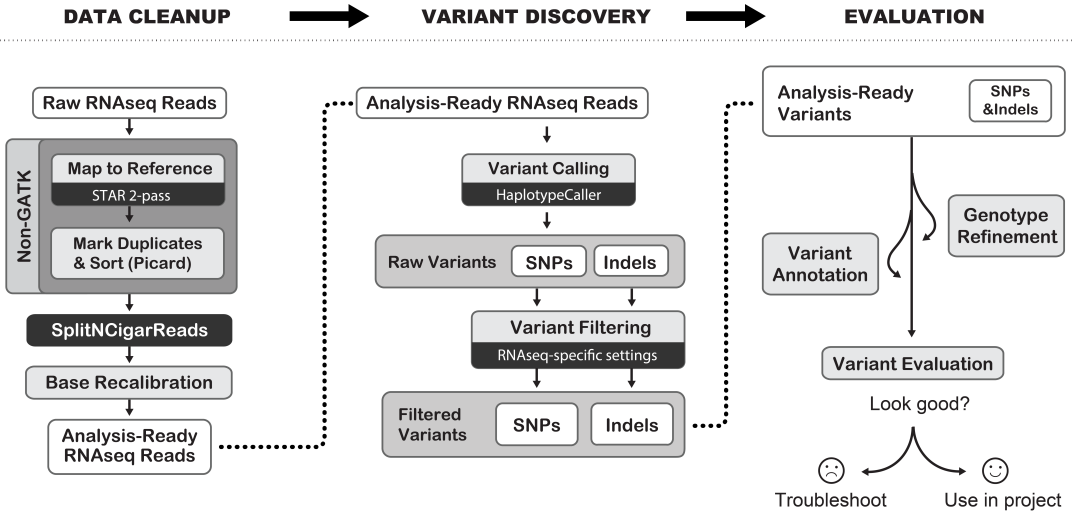
- Generation of reference genome index
- Raw read processing
- Raw read QC
- Mapping to reference genome
- Read counting
- Additional read processing for piRNA discovery
- Post-alignment QC metrics
- piRNA cluster discovery
- Processing of piRNA clusters
- Generation of QC plots
- Aggregation of QC metrics and program logs

Variant discovery in RNA-seq data (GATK)



Source: GATK

Variant discovery in RNA-seq data (GATK)

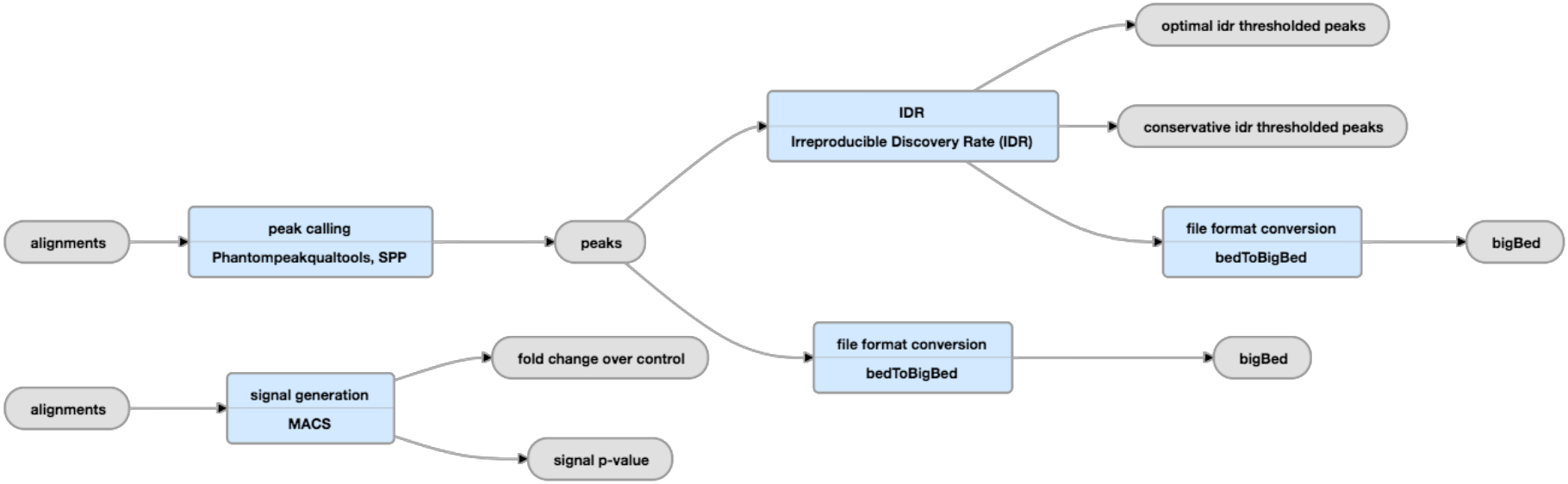


- Mapping to reference genome using a 2-pass protocol to gain accuracy on splice junctions
- Alignment processing: marking duplicates, refinement of raw alignments close to splice sites and base recalibration
- Variant calling
- Variant filtration
- Variant annotation

Source: GATK

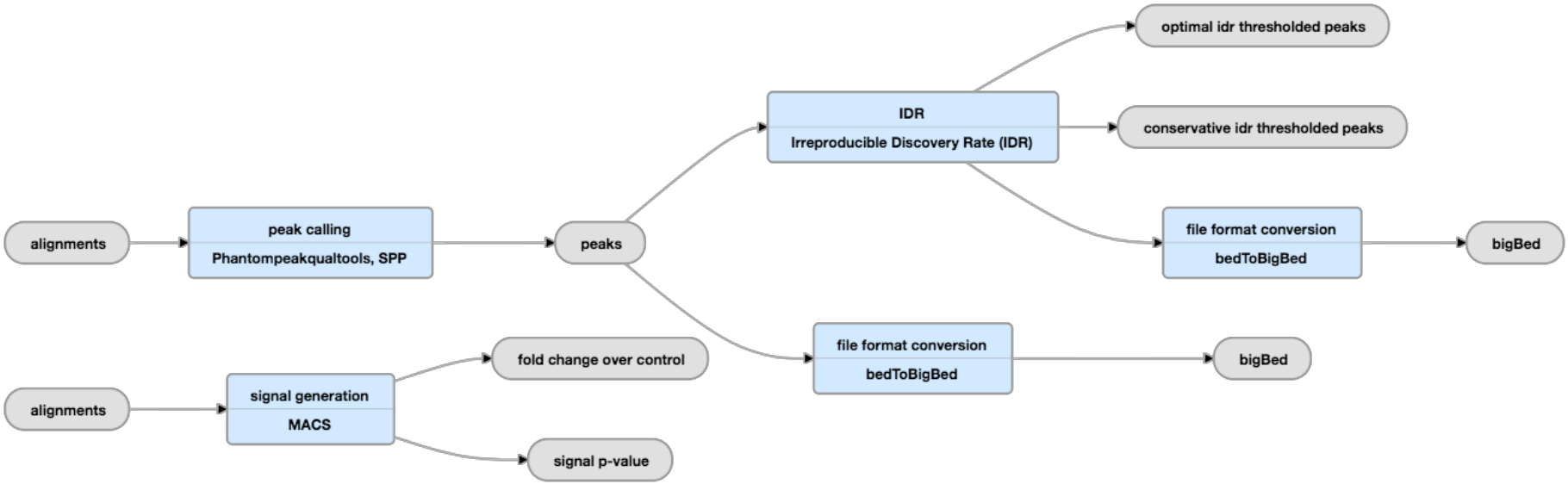
1. Introduction to workflows frequently used in bioinformatics of NGS data
2. Workflows for transcriptomics
    1. Bulk RNA-seq
    2. Small RNA-seq
    3. Variant discovery in RNA-seq data
3. Workflows for functional genomics
    1. ChIP-seq
4. Comments

Workflow for ChIP-seq data processing



Source: ENCODE

Workflow for ChIP-seq data processing

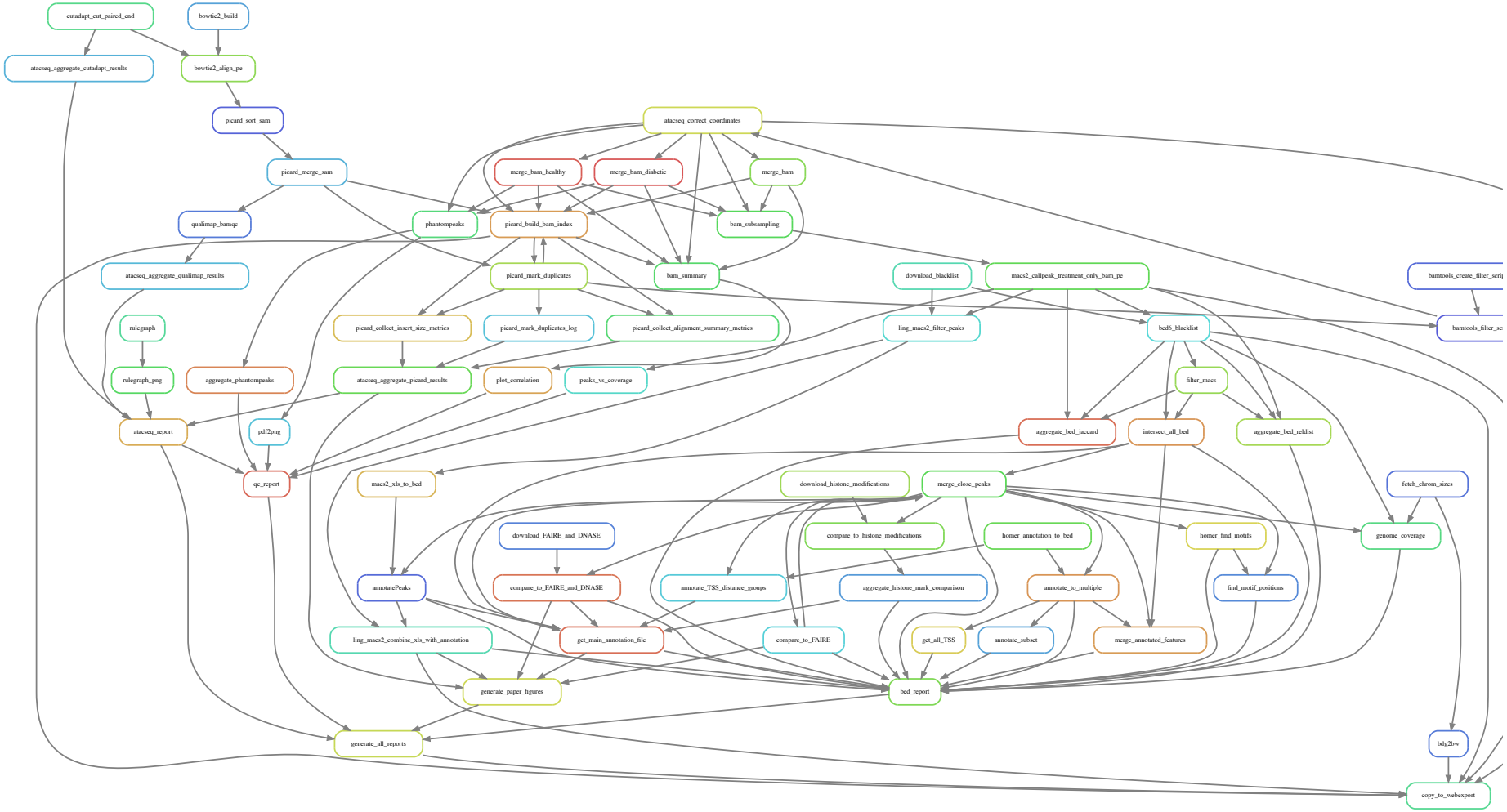


Starting point is processed alignments (already performed: marking duplicates, filtering out reads mapped to blacklisted regions, retaining only reads with one best alignment; QC; fragment length estimation)

- Peak detection
- Peak filtering
- IDR calculation
- Generation of coverage tracks

Source: ENCODE

Workflow for ChIP-seq data processing



A close − to − truth example of a bioinformatics workflow

To show this poll

1

Install the app from pollev.com/app

2

Start the presentation

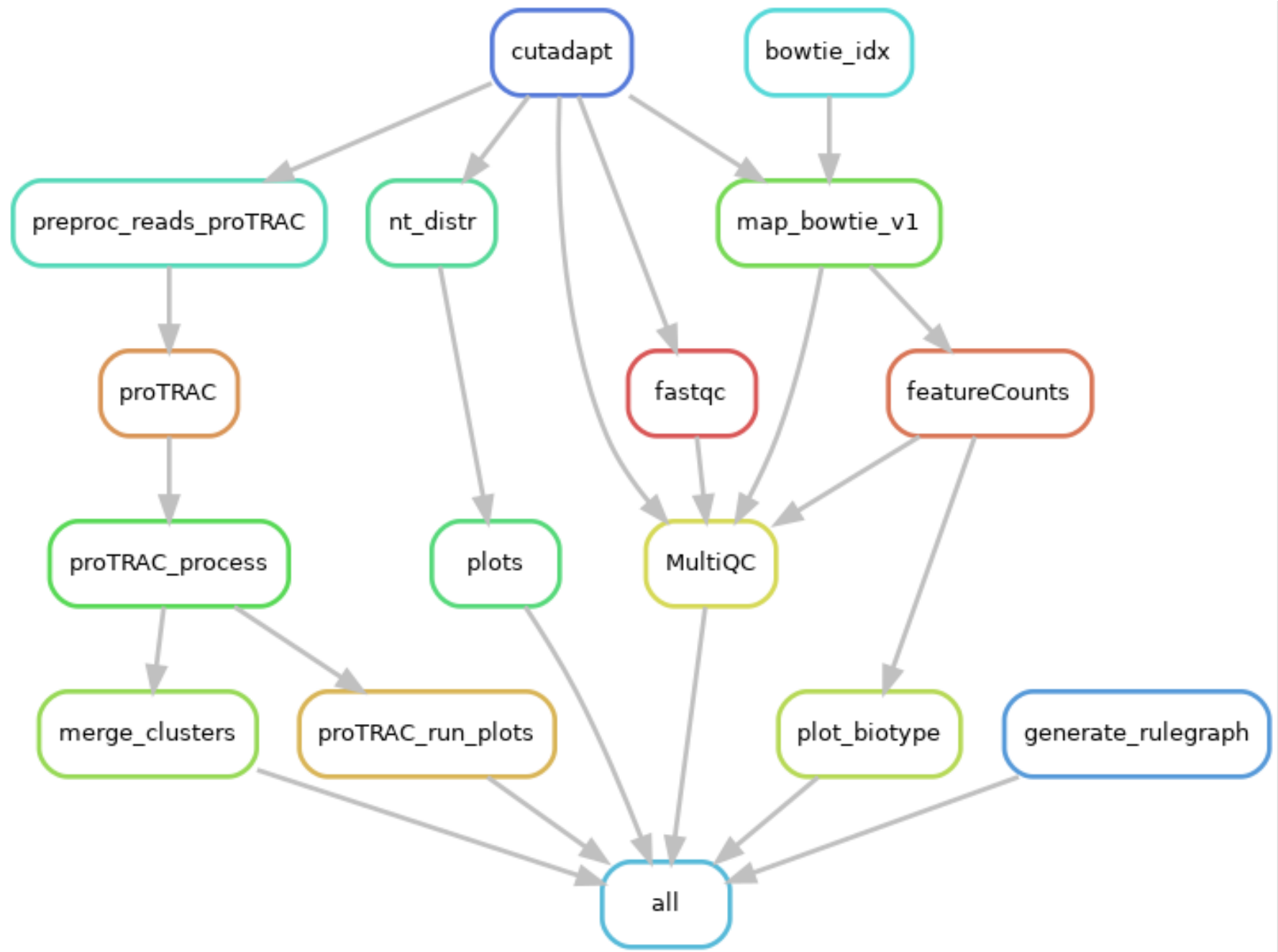
Still not working? Get help at pollev.com/app/help

or

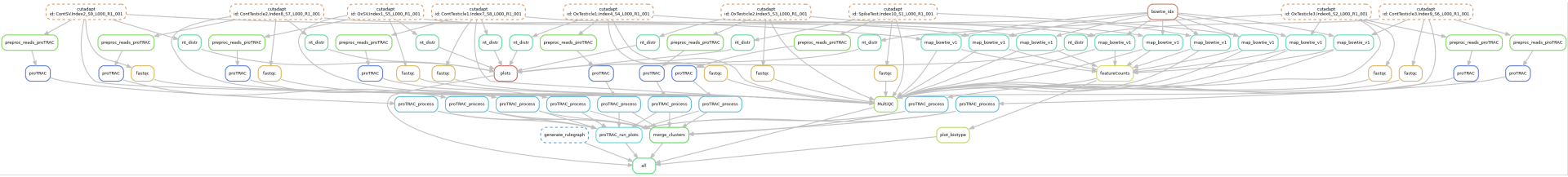
Open poll in your web browser

1. Introduction to workflows frequently used in bioinformatics of NGS data
2. Workflows for transcriptomics
    1. Bulk RNA-seq
    2. Small RNA-seq
    3. Variant discovery in RNA-seq data
3. Workflows for functional genomics
    1. ChIP-seq
4. Comments

Small RNA-seq (piRNA)



piRNA pipeline when fate of each sample is depicted



9 samples

Workflow manager

- Workflow manager is tool for creating **reproducible and scalable** data analyses workflows that consist of a series of processing stages - known as '**pipelines**'.
  - Snakemake
  - NextFlow
  - BPipe
- Advantages: automatic execution of the steps, handling failed jobs, easy restarting, easy parallelisation, compartmentalisation, integration with cluster resource managers, automatic logs, …
- Cons: one has to learn a new tool to run the actual tools.

So if not “One-size-fits-all”, what to do?

- Select only open source solutions and verify whether the workflow suits the needs of your project (tools, versions, reference, parameters).
- Check for updates − is the workflow maintained? Does it support / implement the newest versions of common tools (i.e. read aligners, peak calling tools, QC tools, etc.)? Is the documentation understandable?
- How many other people use it? − Is is a one-project workflow that also has an accompanying separate publication or is it generally accepted in the community?
- Contact an expert to discuss your choice (collaborator, colleague, NBIS consultation).

Thank you

for your attention

Hands-on exercise

- Basic Unix commands to navigate server environments and transfer data
- Clone and use a git repository
- RNA-seq data processing workflow
  - QC of raw reads
  - Alignment format conversions
  - QC of mapped reads
  - Generating QC reports
  - Data inspection in a genome browser